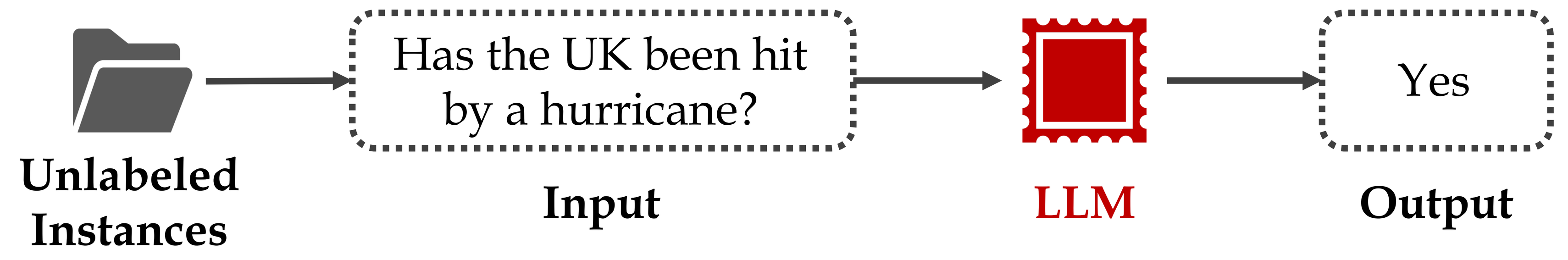
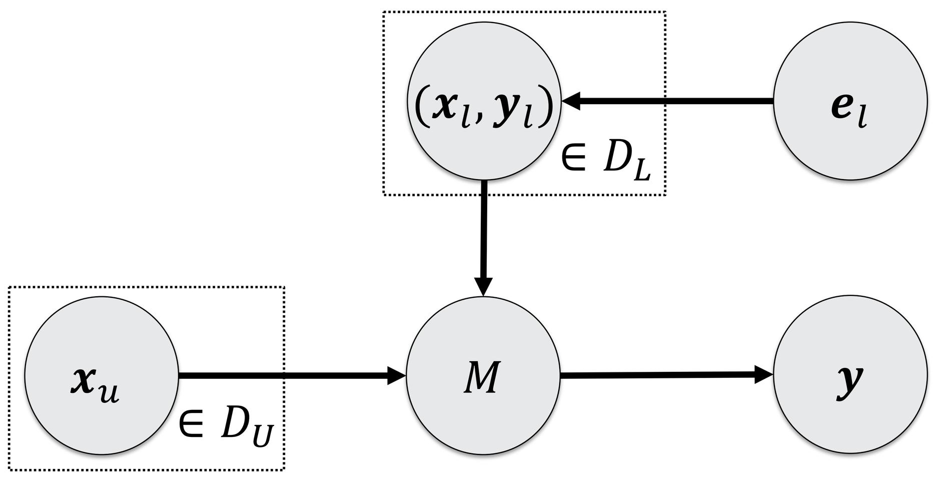
Making Large Language Models Better Data Creators

Dong-Ho Lee[1], Jay Pujara[1], Mohit Sewak[2], Ryen W. White[2], Sujay Kumar Jauhar[2]

[1]Information Science Institute, University of Southern California, [2]Microsoft Research

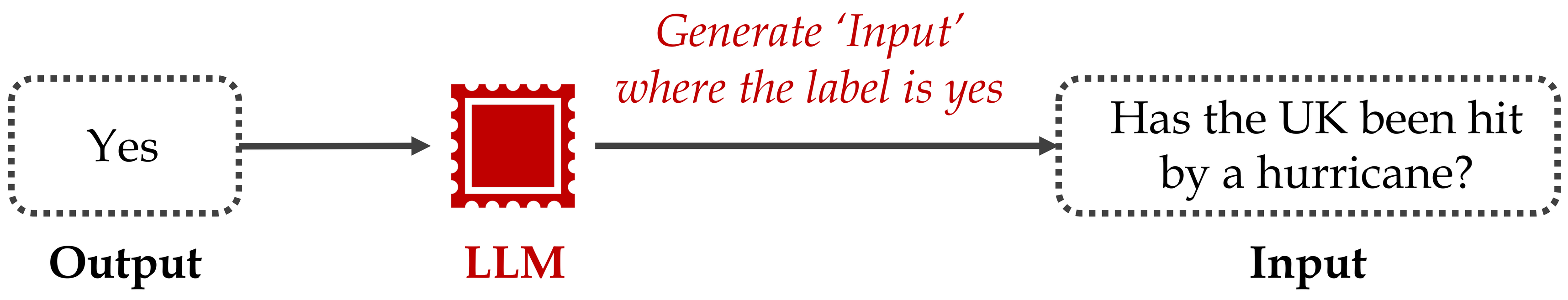
LLM as Labeler

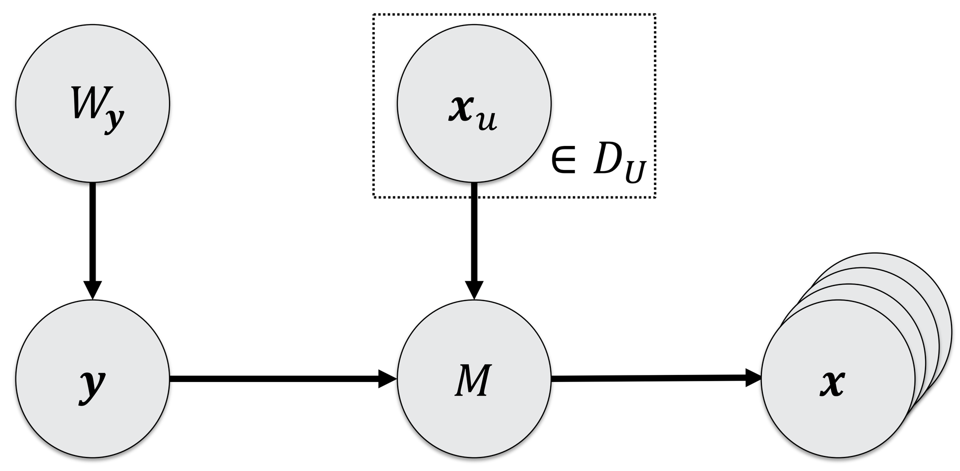




Need unlabeled examples D_U
Curating diverse and representative examples to label is challenging.

LLM as Generator

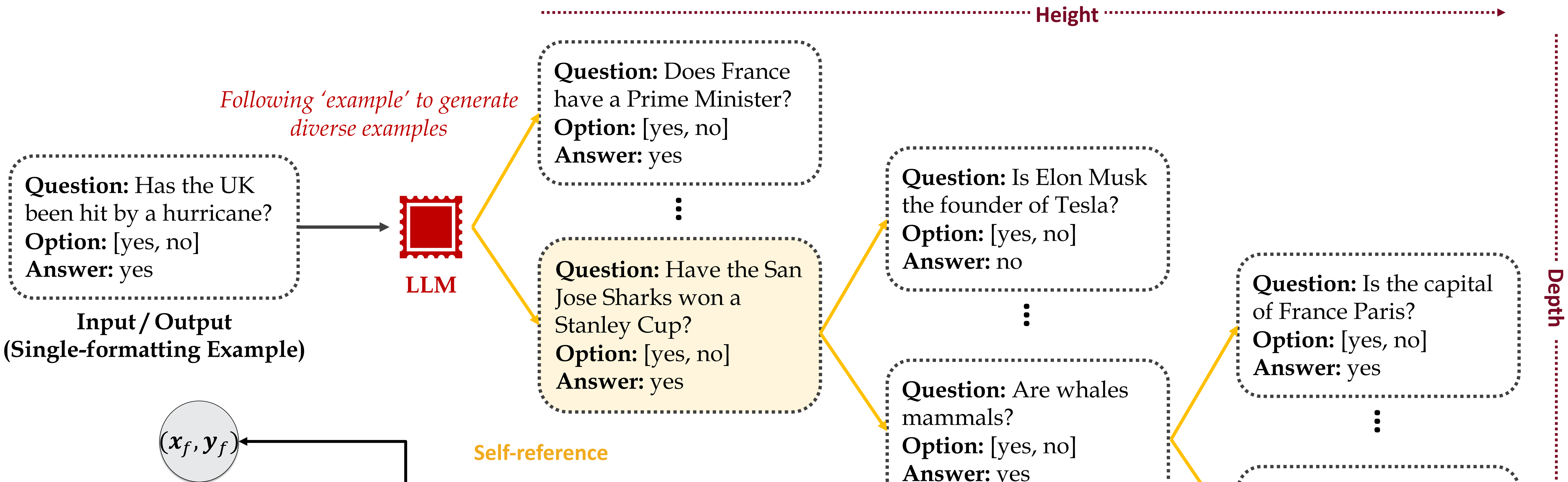


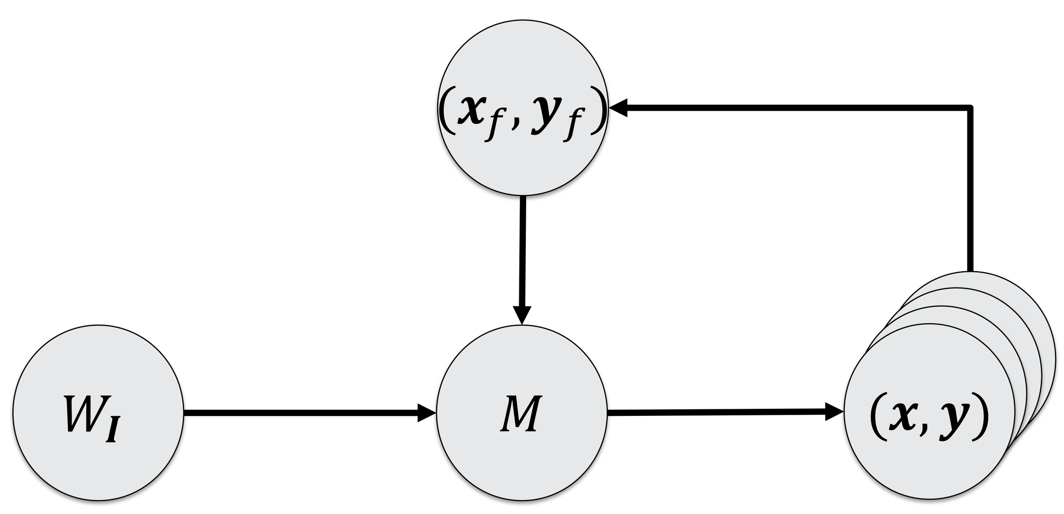


Semantic meaning of y
If y is an index or binary response, it is difficult to generate examples

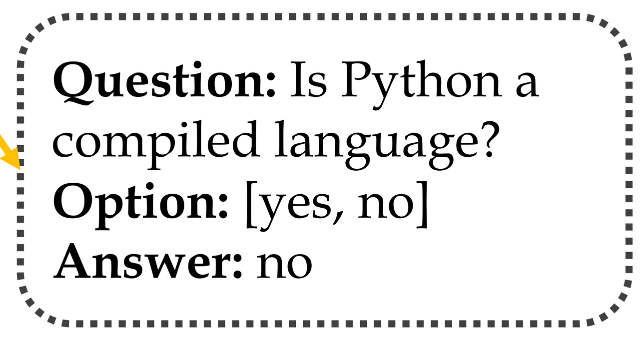
Need specific prompt W_y
If the label is "entailment" for NLI, prompt should include "s1, in other words"

Example-based Data Creation





1) **Random** selection: DFS
2) **Contrastive** selection: DFS by selecting the large sematic different example
3) **Similar** selection: DFS by selecting the large semantic similar example
4) **Tree** selection: BFS



ID Performance

	MCQA (2)		MCQA (5)		Open Yes/No			Closed Yes/No		
Trained on ↓	PIQA	WinoGrande	CommonsenseQA	RiddleSense	BoolQ	PubMedQA	BioASQ	BoolQ	StrategyQA	CREAK
# Examples in $\mathcal{D}$	14,113	160	8,500	3,510	9,427	450	670	9,427	2,061	10,176
$\mathcal{D}_L$	80.95	51.41	68.17	56.48	85.62	55.20	87.14	65.68	49.56	81.19
$\mathcal{D}_G$ (Random)	66.20	51.26	42.06	37.85	68.99	59.80	80.71	52.23	53.04	67.93
$\mathcal{D}_G$ (Contrastive)	66.15	52.36	41.57	38.43	66.66	59.20	67.14	**61.28**	49.56	67.93
$\mathcal{D}_G$ (Similar)	67.15	52.05	47.62	42.09	69.60	60.60	83.57	**61.28**	49.56	69.24
$\mathcal{D}_G$ (Tree)	**68.35**	**52.81**	**48.50**	**42.26**	**69.66**	**61.60**	**85.71**	**61.28**	**56.52**	**72.74**
$(\mathcal{D}_G - \mathcal{D}_L)/\mathcal{D}_L$	−18.43%	+2.65%	−40.55%	−33.64%	−22.91%	+10.38%	−1.66%	−7.18%	+12.31%	−11.61%

- **LLMs can play an important role when access is only available to little data (PubMedQA, BioASQ, WinoGrande)**
- **(Tree)-based exploration limits the semantic distance between the seed sample and the created instances**

API Cost (USD)

Dataset	# Train	Random	Diverse	Similar	Tree
PIQA	14,113	3.60	**2.82**	3.62	3.97
WinoGrande	160	**0.02**	**0.02**	0.03	**0.02**
CommonsenseQA	8,500	2.73	2.71	2.77	**1.73**
RiddleSense	3,510	**0.95**	**0.95**	1.00	1.05
BoolQ	9,427	5.13	**2.24**	4.95	4.2
PUbMedQA	450	0.17	**0.15**	0.17	0.17
BioASQ	670	0.24	0.23	0.33	**0.22**
BoolQ	9,427	3.13	4.10	3.22	**3.11**
StrategyQA	2,061	0.66	0.70	0.81	**0.66**
CREAK	10,176	3.24	**3.20**	4.14	3.50

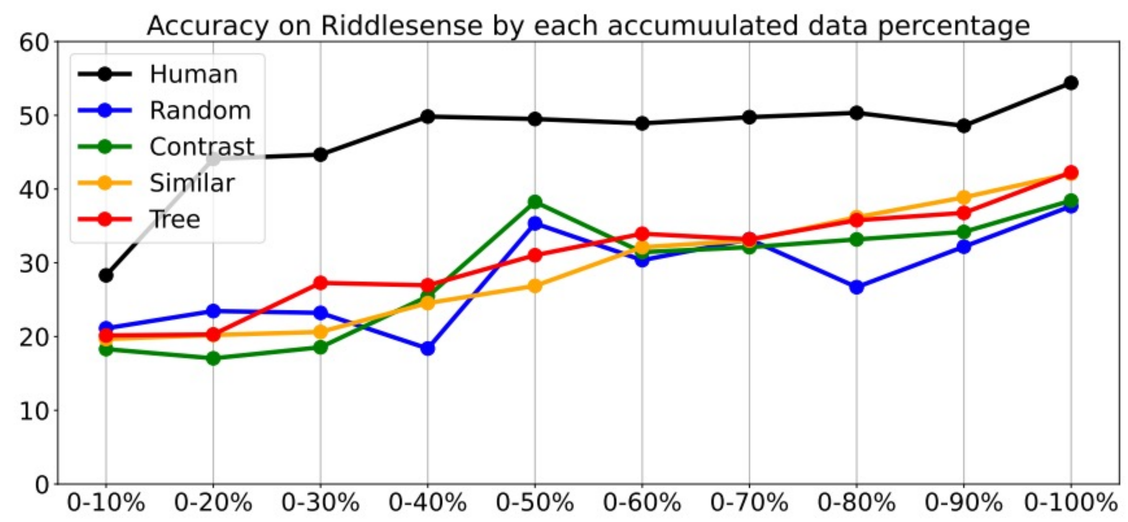
gpt-3.5-turbo as of June 2023
(0.002 USD per 1K tokens)

OOD Performance

	MCQA (2)		MCQA (5)		Open Yes/No				Closed Yes/No	
Train → Trained on ↓ Test →	PIQA WinoGrande	WinoGrande PIQA	CommonsenseQA RiddleSense	RiddleSense CommonsenseQA	BoolQ PubMedQA	PubMedQA BoolQ	BioASQ PubMedQA	PubMedQA BioASQ	StrategyQA CREAK	CREAK StrategyQA
$\mathcal{D}_L$	**52.05**	44.65	41.51	40.93	62.80	58.65	67.14	56.20	49.27	48.69
$\mathcal{D}_G$ (Random)	51.57	49.10	38.51	41.33	59.00	55.77	66.42	59.40	49.27	48.69
$\mathcal{D}_G$ (Contrastive)	50.31	49.50	32.94	42.35	59.00	59.87	75.00	55.20	49.27	46.95
$\mathcal{D}_G$ (Similar)	48.42	**52.25**	**43.42**	42.62	**64.60**	**62.50**	77.85	63.00	49.27	51.30
$\mathcal{D}_G$ (Tree)	50.31	49.55	40.09	**43.35**	**64.60**	61.28	**81.42**	**66.00**	**57.72**	**54.78**
$(\mathcal{D}_G - \mathcal{D}_L)/\mathcal{D}_L$	−0.93%	+14.54%	+4.39%	+5.58%	+2.78%	+6.16%	+17.53%	+14.84%	+14.63%	+11.11%

- **Models trained on LLM data are showing strong generalizability.**
- **Robustness and generalizability of real-world systems that often deal with inputs that are very different from carefully curated academic datasets.**

Size of created data



https://github.com/microsoft/llm-data-creation